Themis: Batch Pre-processing for Ultrahigh Resolution Mass Spectra of Complex Mixtures

Remy Gavard,[*,†] David Rossell,[‡,¶] Simon E.F. Spencer,[‡] and Mark P. Barrow[*,§]

†*MAS CDT, University of Warwick, Coventry, CV4 7AL, United Kingdom*

‡*Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom*

¶*Department of Economics & Business, Universitat Pompeu Fabra, Barcelona, 08005, Spain*

§*Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom*

E-mail: R.Gavard@warwick.ac.uk; M.P.Barrow@warwick.ac.uk

Phone: +44 (0) 24 76151013. Fax: +44 (0) 24 76524112

Abstract

Fourier transform ion cyclotron resonance mass spectrometry affords the resolving power to determine an unprecedented number of components in complex mixtures, such as petroleum. The software tools required to also analyze these data struggle to keep pace with advancing instrument capabilities and increasing quantities of data, particularly in terms of combining information efficiently across multiple replicates. Improved confidence in data and the use of replicates is particularly important where strategic decisions will be based upon the analysis. We present a new algorithm named Themis, developed using R, to jointly preprocess replicate measurements of a sample with the aim of improving consistency as a preliminary step to assigning peaks to chemical compositions. The main features of the algorithm are quality control criteria to detect failed runs, ensuring comparable magnitudes across replicates, peak alignment

and the use of an adaptive mixture model-based strategy to help distinguish true peaks from noise. The algorithm outputs a list of peaks reliably observed across replicates and facilitates data handling by preprocessing all replicates in a single step. The processed data produced by our algorithm can subsequently be analyzed using relevant specialized software. While Themis has been demonstrated using petroleum as an example of a complex mixture, its basic framework will be useful for complex samples arising from a variety of other applications.

Introduction

Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS)[1–6] represents a state-of-the-art technique for the study of complex mixtures that provides significant advantages in terms of ultrahigh resolving power and mass accuracy[7] . As a result of these performance advantages, FTICR MS affords the ability to distinguish molecules with very similar mass-to-charge ratios (m/z), on the basis of mass defect. Given the complexity of petroleum composition, these advantages are particularly relevant for the characterization of petroleum and its products by mass spectrometry[8–14] , an area of research that has become known as "petroleomics" . The following discussion will use application to this field as a suitable example, but it should be made clear that our methodology remains applicable to other complex samples. A variety of analytical approaches have been applied for the characterization of petroleum[15] , as well as environmental samples associated with alternative sources of oil[16–19] . Although high-field Orbitrap mass spectrometers are showing promising results for light and medium petroleum fractions, FTICR MS remains state-of-the-art for heavy fractions[20–24] . In order to address the challenges of producing and refining crude oil, one needs to develop a more detailed understanding of its composition through improvements in characterization methods[25,26] . Petroleomics is a field of growing importance because the most desirable varieties of crude oil are becoming more scarce. At the same time, the derivatives of crude oil are in everyday use and include products such as fuels, solvents, plastics,

dyes, waxes, lubricants, and pharmaceuticals, among others[27] .

As the capabilities of FTICR MS have increased and produce larger and richer datasets, there has been an accompanying need for the development of more advanced software for data analysis[28] . Peak detection is a fundamental step as part of a data analysis workflow, regardless of application and instrument type. Reflecting this, a large variety of methods have been developed over time to improve peak picking[29–37] . Thus far, the development of data analysis methodologies for mass spectrometry have focused upon the characterization of biomolecules, such as peptides and proteins. In 2003 , Patterson argued in relation to the study of biomolecules that 'data analysis is the Achilles heel of proteomics and our ability to generate data now outstrips our ability to analyze it'[38] . Today, the ability to analyze proteomics data is considerably improved, with many software tools available. The analysis of complex mixtures data[29–31,39–42] is different from proteomics, metabolomics, or polymer data, for example, given the higher peak density (15 to 30 peaks in a 0.5 m/z window)[10,12,35] and different patterns within the data. While proteomics has typically involved lower resolution instrumentation and higher throughput techniques (automated system analyzing many samples per day), of greatest need when analysis petroleomic samples is ultra-high resolution, making FTICR MS the tool of choice. Another difference is that software tools for biomolecule characterization are designed to match protein or peptide sequences using online databanks. For complex mixtures such as petroleum, the strategy is to determine series of heteroatom containing organic components, with thousands of possible compositions ($C_cH_hN_nO_oS_s$).

One example of data analysis software is Mass-Up[43,44] , an open source mass spectrometry program that gathers functions such as normalization, peak detection and peak matching of replicated samples. It was developed specifically for proteomics MALDI data[45–47] , when typically using a lower resolution mass analyzer such as time of flight mass spectrometry. While a software tool designed for other varieties of mass analyzers and other sample types can be invaluable for their intended purposes, they are not appropriate for complex mixtures

analysis due to their design for use with lower resolution data and wider mass error tolerances (*e.g.*: hundreds of *parts per million*, *ppm*). There is an emerging need for improved data analysis strategies for complex mixtures, such as for petroleomics applications, that are designed for the resolution of ten of thousands of peaks[15] .

Currently, a typical workflow for analysis using FTICR MS may consist of acquiring one spectrum per sample and processing each individual sample with specialized petroleomics software, such as Composer (Sierra Analytics, Modesto, CA, U.S.A.)[16,20] or PetroOrg (Florida State University, Tallahassee, FL, U.S.A.)[48] . The results from individual samples can then be recalibrated with respect to m/z to compensate for electric field effects (including space-charge due to the presence of the ions) within FTICR cells[13,49–51] . As the field becomes more mature, increasing numbers of samples need to be analyzed within a practical time frame, including multiple experiments to ensure repeatability of results. A fundamental concern is to ensure that the data are reliable and false assignments are reduced by removing as much noise as possible before performing in-depth data analysis.[36] .

To improve the reliability of analysis of crude oil spectra, Hur et al. have previously highlighted the importance of the use of replicates[52,53] . The need for replicates was demonstrated for FTICR MS-based metabolomics data[54] , and recently replicates were used to generate an averaged mass spectrum[55] . Our approach is based on the idea that to fully capitalize on the advantages brought by repeat measurements, replicates should be processed together instead of separately. The first challenge is that complex mixture datasets present a high density of peaks of interest, hampering the identification of those that are consistent across replicates. A second challenge is that of the peak magnitudes: some peaks are similar in magnitude to the noise level, and it is also possible that peak magnitudes can differ significantly across replicates.

A simple strategy to avoid false positives is to use stringent parameters when making peak assignments, *e.g.* setting a higher minimal signal-to-noise (S/N) ratio when picking peaks or a narrower tolerance of mass error (more limited deviation on the m/z axis). There

are advantages in working with such peak lists rather than full mass spectra in terms of simplicity and reduced computational cost. The problem with these strategies is that they may, at an early stage, discard low magnitude peaks that provide valuable information and are consistently observed across replicates. That is, they may be too aggressive in reducing the number of peaks, with consequences for subsequent interpretation. In contrast, using settings that are too permissive risks including a high number of false positives. Further, the fundamental issue remains that applying thresholds to individual spectra loses the opportunity to share information across samples. Ideally, one would like to preserve all potential peaks in individual samples and then use information across replicates to identify which peaks are truly reliable. Traditionally, denoising methods are based on signal magnitude, either using the shape of the peaks or their magnitudes, to discriminate between noise and reliable peaks. By contrast, we propose to denoise the spectra by focusing upon the consistency on the m/z scale, with peak magnitude being used as a secondary criterion. Our algorithm ensures reproducibility of the peak list extracted from a sample and produces a single consensus list. Figure 1 provides a schematic representation. The first stage is to extract a peak list from each replicate using a permissive S/N ratio. The second step is to detect anomalous replicates using quality control statistics based upon their molecular weight distributions. The third stage is the use of quantile normalization to ensure that magnitudes are comparable across replicates. Finally the fourth step uses a statistical mixture modeling approach to distinguish reliable peaks from those due to noise.

Methodology

Sample Preparation

Sample A was an NIST light sour crude oil sample (National Institute of Standards and Technology, SRM 2721, Crude Oil (Light -Sour)) which was dissolved as 0.1 mg/mL in an 80:20 ratio of propan-2-ol/toluene (Fisher Scientific, Loughborough, UK), with formic

acid (Sigma-Aldrich Company Ltd, Gillingham, UK) being added as 1 % by volume to aid protonation. Sample B was a South American crude oil sample which was dissolved as 0.05 mg/mL in a 50:50 propan-2-ol/toluene (Fisher Scientific, Loughborough, UK) and with 0.2% formic acid (Sigma-Aldrich Company Ltd, Gillingham, UK) for positive-ion mode or ammonium hydroxide (Sigma-Aldrich Company Ltd, Gillingham, UK) at 0.8% for negative-ion mode. Sample C was a Kodak naphthenic acid (NA) mixture (The Eastman Kodak Company, Rochester, NY) was prepared at 0.1 mg/mL in acetonitrile (VWR Chemicals, Lutterworth, UK) without the addition of any ammonium hydroxide.

Instrumentation

Mass spectra were acquired using an Apollo II electrospray ionization (ESI) source, coupled to a 12 T solariX FTICR mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany). For sample A, the instrument was operated in positive-ion mode and 6 repeat measurements were obtained all of them being the result of 300 scans. Sample B was recorded in both positive and negative mode with 5 and 6 repeat measurements, respectively. The number of scans was 300 for the negative mode and 210 scans for the positive mode. Sample C was recorded in negative mode with 6 repeat measurements and 100 scans. In all cases, replicates were obtained the same day using a single session on the instrument. Broadband mass spectra were acquired, where a single zero fill and Sine-Bell apodization were applied before usage of a Fourier transform.

Statistical processing

Extract peak lists: The spectra were exported from solariXcontrol to DataAnalysis 4.2, the latter was used to extract peak information using the following parameters: Peak finder "FTMS", S/N threshold of 4, relative magnitude threshold (base peak) of 0.01 % and absolute magnitude threshold of 100%. The spectrum was not subject to any modification other than the application of the default apodization before undergoing the Fourier transformation.

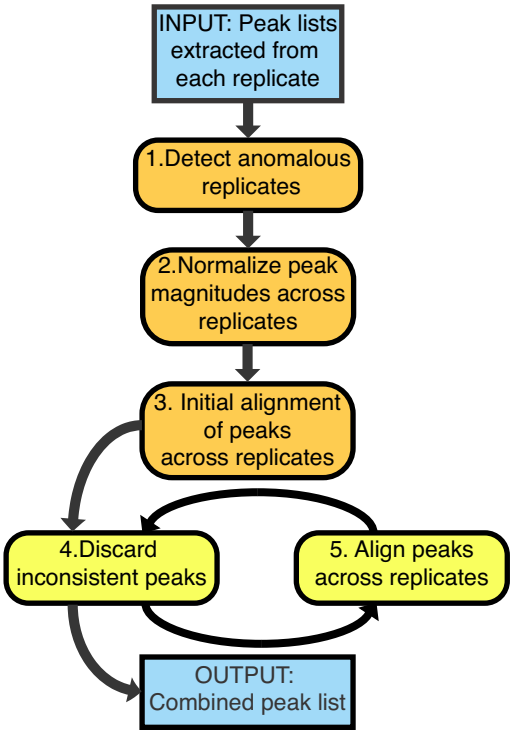

Figure 1: Schematic of the Themis pre-processing algorithm

Step 1: Detect anomalous replicates. The average molecular weight $\overline{W}_j$ of each replicate $j = 1, \ldots, r$, where r is the number of replicates, was calculated as a quality control metric to detect anomalous runs. Specifically

$$\overline{W}_j = \frac{\sum_{i=1}^{n_j} M\left[i,j\right] I\left[i,j\right]}{\sum_{i=1}^{n_j} I\left[i,j\right]} \tag{1}$$

where $M[i,j]$ is the m/z value of peak i in the sample j, $I[i,j]$ is the corresponding magnitude and n_j is the number of peaks in sample j.

To identify what constitutes an anomalous average molecular weight, we must first characterize their reference distribution from the data. Given that the mean and the standard deviation (sd) can be heavily influenced by outliers, we used robust measures of the center and spread, namely the median and the corrected median absolute deviation (mad) [56–58] given by:

$$\text{mad}(x_1, \ldots, x_n) = b\left(\text{median}_i(|x_i - \text{median}_j(x_j)|)\right) \tag{2}$$

with $b = 1.4826$ for Gaussian distributions. Motivated by the Central Limit Theorem, we assume that the average molecular weights of non-anomalous samples are approximately normally distributed around a mean μ, with standard deviation σ. We wish to find an interval $(\mu - y,\ \mu + y)$ that in the absence of any anomalies should contain all n samples with probability $1 - \alpha$, where α is a user-specified error threshold (by default $\alpha = 0.05$). Assuming that replicates are independent, for a given μ and σ it can be seen that:

$$y = \Phi^{-1}\left(\frac{(1-\alpha)^{1/r}}{2}, \mu, \sigma\right) \tag{3}$$

where $\Phi^{-1}(x, \mu, \sigma)$ is the inverse normal cumulative distribution function.

Step 2: Normalize peak magnitudes across replicates. To take into account that the dynamic range of magnitudes varies across samples we apply quantile normalization [59,60] . This ensures that the distribution of magnitudes is identical across replicates, facilitating the subsequent peak alignment.

Step 3: Initial alignment of peaks across replicates. Our peak alignment strategy has two steps, a first one to initialize (Step **3**) and a second one used iteratively to refine the matching Step (**5**) . For clarity we denote any value that may change across iterations with a k superscript to indicate the value at the kth iteration. In the initialization step $k = 0$. To initialize the peak alignment, we take the sample with the largest number of peaks as a reference and match peaks in all the other replicates to the reference. Let $m^{(k)}$ denote the number of aligned peaks in iteration k and $m^{(0)}$ the number of peaks in the longest replicate at initialization. For each peak in the reference replicate we match to the closest peak in each replicate in terms of its m/z value.

Step 4: Discarding inconsistent peaks. We compute the standard deviation of the m/z values matched to reference peak $i = 1, \ldots, m^{(k)}$, which we denote $Z_i^{(k)}$. Intuitively, peaks that are consistently observed across samples should show similar m/z values, resulting in low $Z_i^{(k)}$. That is, one typically observes a sub-population of reliable peaks with low $Z_i^{(k)}$

and another sub-population of less reliable peaks with high $Z_i^{(k)}$, likely due to noise. This motivated us to fit a mixture model to separate these subpopulations. Let $P_{ij}^{(k)} \in \{1, \ldots, n_j\}$ be the index of the peak in replicate j (for $j = 1, \ldots, r$) that is matched to the i^{th} reference peak in iteration k. We define the mean m/z and magnitude for reference peak $i = 1, \ldots, m^{(k)}$ by

$$\overline{M}_i^{(k)} = \frac{1}{r} \sum_{j=1}^{r} M[P_{ij}^{(k)}, j] \tag{4}$$

$$\overline{I}_i^{(k)} = \frac{1}{r} \sum_{j=1}^{r} I[P_{ij}^{(k)}, j] \tag{5}$$

and

$$Z_i^{(k)} = \sqrt{\frac{\sum_{j=1}^{r} \left(M[P_{ij}^{(k)}, j] - \overline{M}_i^{(k)} \right)^2}{r-1}} \tag{6}$$

$$T_i^{(k)} = \sqrt{\frac{\sum_{j=1}^{r} \left(I[P_{ij}^{(k)}, j] - \overline{I}_i^{(k)} \right)^2}{r-1}} \tag{7}$$

the respective m/z and magnitude standard deviations.

An important step in our algorithm is to identify the subpopulation of peaks consistently observed across replicates. To this end we fit a Normal mixture model[61] to $\log\left(\frac{Z_i^{(k)}}{\overline{M}_i^{(k)}}\right)$ using the function `mclust`[62,63] in the R package mclust. Calculating the relative standard deviation (RSD) by dividing the standard deviation $Z_i^{(k)}$ of a peak by its $\overline{M}_i^{(k)}$ allows us to express the results in a unit equivalent to *parts per million* which is a standard unit when expressing the mass error associated with the m/z of a peak. In addition, it helps to make the mixture model more reliable as it allows to be equally stringent for high and low m/z as the sd will naturally be higher with high m/z values. We denote $G_i^{(k)} = \log\left(\frac{Z_i^{(k)}}{\overline{M}_i^{(k)}}\right)$.

In `mclust`, we set the maximum number of components to capture peak subpopulations of high, low and potentially a third one of intermediate quality. We use the Bayesian information criterion (BIC) to select the final number of components in the mixture. Themis then selects the population with lowest mean $G_i^{(k)}$. When this mean is $> 1ppm$ a warning is given

to signal that the dataset may be of low quality. The first time that Step **4** is performed, a conservative threshold is used: peaks are discarded if they have a probability below 0.01 of belonging to the selected subpopulation. Doing so allows the algorithm to remove the majority of the obvious noise whilst making sure not to discard any potentially relevant peaks. At this step, the presence of leftover noise isn't problematic as further refining will be performed by iteratively repeating Steps **4** and **5** .

In each subsequent repetition of Step **4** in future iterations the 0.01 threshold is increased by 0.01, up to a maximum of 0.5. The goal is that by the end of the iterative process only peaks belonging to the high-quality subpopulation remain.

Step 5: Align peaks across replicates. After peak removal in Step **4**, we refine the peak matching across samples using a combined criterion that incorporates both magnitude and m/z, in contrast to Step **3** where we only used m/z. Intuitively, the criterion seeks the closest peak based on a score where m/z and magnitude are weighted according to their inherent variability. Given that the precision of the variance estimates in (6)-(7) may suffer when the number of replicates r is low, we borrow strength across peaks using the hierarchical empirical Bayes framework proposed by Smyth and Speed[64] , implemented in function `squeezeVar` from the Bioconductor package limma[65] . We denote $\tilde{Z}_i^{(k-1)}$ and $\tilde{T}_i^{(k-1)}$ the refined estimates analogous to Z_i^{k-1} and T_i^{k-1}. Specifically, the score to measure the closeness of peak ℓ in sample j to reference peak i at the kth iteration is

$$S_{ij\ell}^{(k)} = \frac{\left|M_{\ell j} - \overline{M}_i^{(k-1)}\right|}{\tilde{Z}_i^{(k-1)}} + \frac{\left|I_{\ell j} - \overline{I}_i^{(k-1)}\right|}{\tilde{T}_i^{(k-1)}}. \tag{8}$$

The highest scoring peak in each replicate replaces the one chosen in the initial matching. After this peak assignment we update $\overline{M}_i^{(k)}, \overline{I}_i^{(k)}, Z_i^{(k)}, T_i^{(k)} \tilde{Z}_i^{(k)}$ and $\tilde{T}_i^{(k)}$. To obtain a scoring method that limits the effect of outliers and can be computed in cases where a reference peak is absent from one or a few replicates, we added the possibility to replace (4)-(5)-(6)-(7) by trimmed means and standard deviations. That is, the replicate(s) with largest $S_{ij\ell}^{(k)}$ in (8)

can be discarded.

Themis iteratively repeats Steps **4** and **5** until either the BIC selects a single population or else when all remaining subpopulations have a mean less than $\equiv 1ppm$ and the peak list does not change between 5 successive iterations.

Output combined peak list. The final output is a list composed of 3 tables containing respectively the m/z values, magnitude values and the final peak list. The m/z and magnitude tables have a $[m^K, r]$ dimension where m^K indicates a peak and r a replicate number. The final reference peak list file is an $m^{(K)} \times 4$ table where $m^{(K)}$ is the number of reference peaks at the final iteration K. Themis stores the m/z, sd(m/z), magnitude and sd (magnitude) of each peak as separated columns. Themis provides a function to extract columns 1 and 3 from the peak list table to a `.txt` file containing a first column with the m/z and a second with the corresponding magnitudes.

Results and discussion

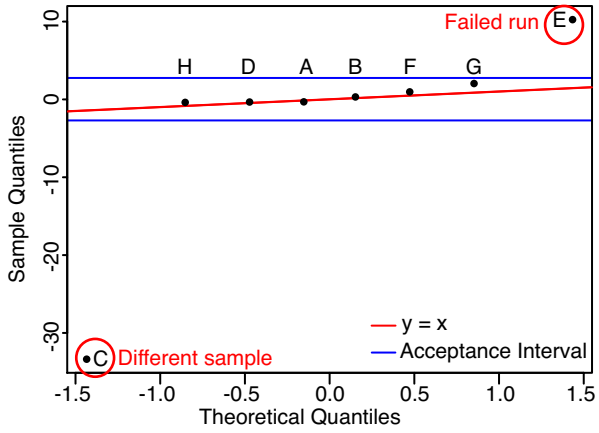

Figure 2: Automated detection of outliers and use of a series of repeat measurements to produce an averaged data set for characterization.

The performance of the pre-processing methodology was assessed using a sample of NIST light sour crude oil, a naphthenic acid sample[66] and a crude oil sample analyzed using both positive-ion and negative-ion modes. We also used a dataset that was recorded using

deliberately aberrant instrument parameters to study the ability of our framework to detect such situations. Themis is available as an online tool at http://themis.warwick.ac.uk/themis and is based on Rwui[67] to generate a web interface for the R script.

A common strategy to improve the accuracy in the m/z values is to apply a calibration step based upon a list of reference peaks. This step can in principle be applied to each individual peak list given as input to Themis or to the single reference peak list output by Themis. It is common that there can be minor variations in mass errors between different datasets. Calibrating each individual peak list before passing to Themis can significantly improve the quality of the processing due to improved consistency.

In order to test Step **1** of the algorithm, we recorded a spectrum of the NIST Sample A, where the ICR cell was intentionally overloaded with a high ion population and one where we deactivated ion source dissociation (ISD), which is used to minimize non-covalent aggregation. These two peak lists were extracted and included with the 6 others which were acquired under normal conditions. The algorithm was able to detect these 2 spectra as aberrant and remove them. Similarly, we then substituted the ISD off peak list by one from naphthenic acid sample C to the list of replicates for sample A (NIST) used before to verify that our method would be able to cover this potential error. Again, the algorithm successfully detected the spectrum which did not correspond to Sample A (NIST) and removed it. The procedure was illustrated in Figure 2 where spectra C and E were discarded after modelisation while the other were kept.

To assess Step **2**, we produced a quantile-quantile plot (q-q plot) to compare the magnitudes across samples. We observed considerable variation between replicates (Figure 3a.), particularly for greater magnitudes. Low magnitudes (ranging from $0 - 0.5 \times 10^8$) exhibited a similar distribution across replicates. In the region from 0.5×10^8 to 2.0×10^8, we observe an inflexion of the line, which demonstrates that the magnitude of the magnitude is different but the overall shape is similar. Also, single high magnitude peaks such as those originating from contaminants will influence the total signal magnitude for the corresponding dataset.

The quantile normalized magnitudes are shown in Figure 3b. Similar results were observed for other samples (see Supporting Information).

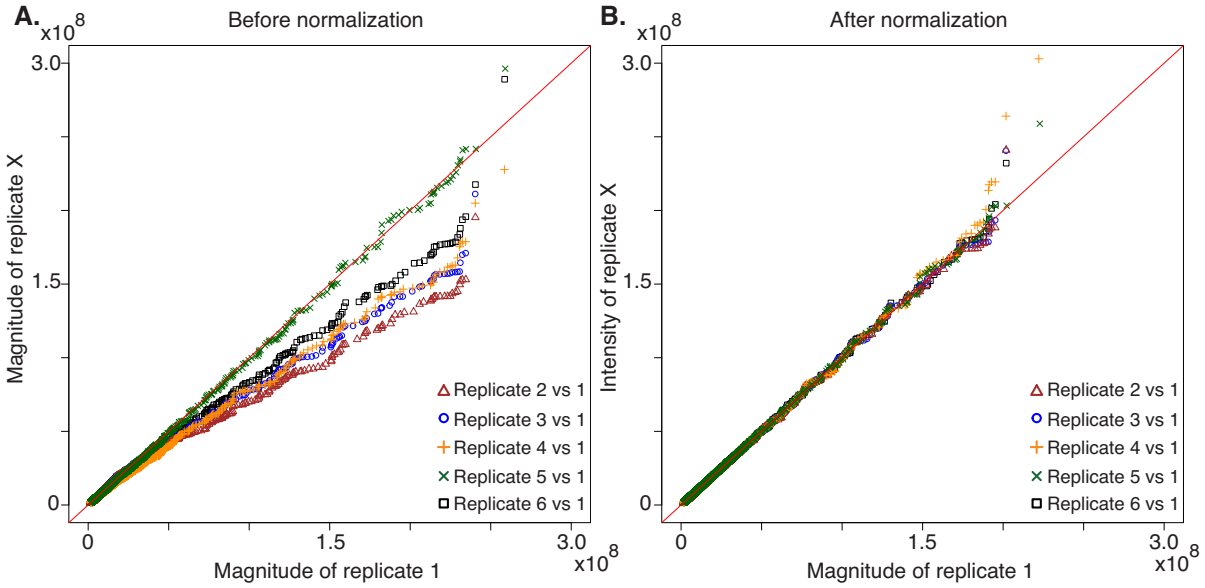

Figure 3: (Quantile-quantile plots of the magnitudes of the six replicates before (**A.**) and after (**B.**) quantile normalization for the NIST light sour crude oil sample.

Figure 4 was produced after the initial peak alignment in Step **3**. It shows a histogram of log-standard deviation within-peak m/z values for multiple datasets. It reveals the presence of a sub-population with low log(sd/mz) corresponding to reliable peaks, *i.e.* with similar m/z across replicates and another sub-population with high log(sd/mz) mostly composed of noise. Evidence of distinct sub-populations was observed in all datasets we have analyzed so far, including different samples, instruments, users and peak list extraction methods.

Step **4** is critical because although in all datasets there are clearly distinct sub-populations, the distributions are different. That is, the threshold used to distinguish reliable from unreliable peaks cannot be a fixed quantity but instead needs to be data-dependent. The red line, labeled "1 *ppm*", indicates a fixed threshold equivalent to the log of 1 *ppm*, a value which is typically used as a benchmark for the accuracy of the mass measurement. For comparison, the black line, labeled "Themis Threshold", indicates the final threshold identified by our

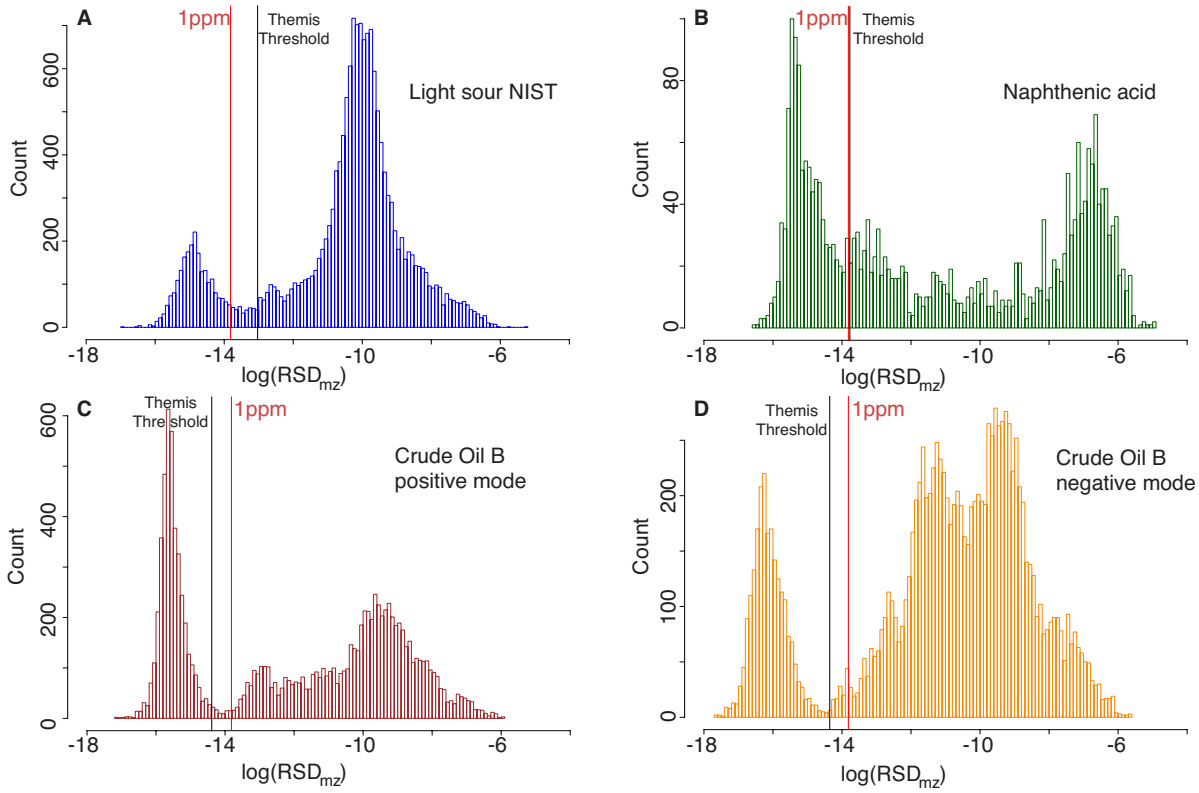

Figure 4: Histograms of the log of the absolute relative standard deviation (RSD) for peaks matched under the initial matching of different samples, the red line, labeled "1 *ppm*", represents a standard deviation equivalent to 1 *ppm*, the black line, labeled "Themis Threshold", the position of the threshold between noise and consistent peaks after Themis processing. **A** is NIST light sour crude oil Sample A, positive-ion ESI; **B** naphthenic acid Sample C, negative-ion ESI; **C** South American crude oil Sample B, positive-ion ESI; **D** South American crude oil Sample B, negative-ion ESI.

mixture model framework, which is adaptive to the nature of the individual datasets. For instance, for both ionization of the Crude Oil B, a more tolerant threshold was used. While for Figure 4C, the threshold immediately makes sense to the eye, Figure 4D may give the impression of selecting part of the noise population. This is because during the refining process the shape of the population changes due to the scoring algorithm. With the NIST data the refinement led to the removal of peaks which ended up being present several times following the rematching performed during the iterative part of the algorithm. During this part the peaks in between the two large populations resulting from valid peaks [on the left] and noise peaks [on the right], noise slowly joined these large populations. The more challenging naphthenic acid sample ended up with a threshold close to 1 *ppm*. This dataset had considerably fewer peaks than the 3 other datasets, making the mixture modeling more challenging. Despite fewer peaks for the mixture modeling, the algorithm still managed to isolate a consistent population.

An example that highlights the benefits of the algorithm is given in Figure 5 with the close examination of a region around the peak $m/z = 248.1434$, for 2 replicate datasets and Themis output using all replicates. It is possible for a user to manually go through every dataset, adjust the parameters, to get an optimal assignment. This is a laborious task which is usually avoided by using default data analysis parameters across the datasets. Manual adjustments of the parameters on a case by case basis is the way to assign the greatest possible number of peaks, but also leads to an increased risk of false assignments due inclusion of noise peaks. In Figure 5 noise was observed between m/z 248.00 and 248.40 for the individual replicates, but was not observed in the dataset produced by Themis.

Figure S5 shows a larger m/z region to illustrate the peak list obtained across the 6 replicates of the NIST sample. Our algorithm identified peaks that were consistently observed across replicates with a S/N ratio as low as 4.5 up to 15 for this section between 700 and 710 m/z. For comparison, in the region around 400 m/z the peaks are routinely observed with a S/N ratio of more than 500.

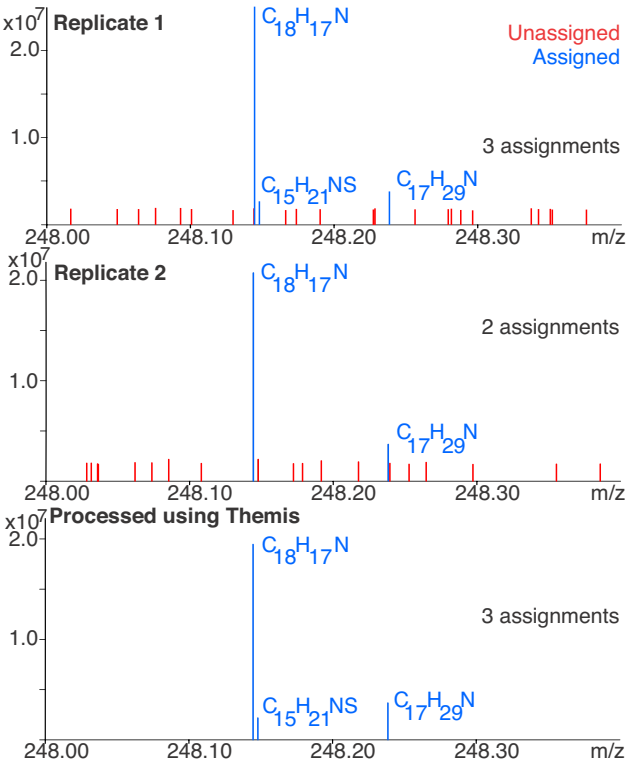

Figure 5: Peak assignment between m/z 248.00 and 248.40, showing 2 replicates datasets and the one produced by Themis using all replicates. In replicate 1 composition $C_{15}H_{21}NS$ is present just above the noise threshold while in replicate 2 the peak is below the noise threshold and so not assigned.

The raw peak lists for the NIST light sour crude oil sample contained an average of approximately 16,400 peaks. Out of these, Themis identified 2,260 reference peaks deemed to be common amongst all of the replicates. The number of entries increased to 2,523 when the peaks were allowed to be absent from one of the replicates at Step **5** of our algorithm, and to 2,820 when peaks could be absent from 2 of the replicates. Allowing peaks to be absent from one or more replicates increases the ability to detect potentially relevant peaks, at the expense of an increased risk of potentially including less reliable peaks.

We compared the chemical composition obtained from the unprocessed spectra with that from the peaks list produced by our algorithm for the NIST light sour crude oil. For the purposes of the comparison, the N_1 class has been used, due to being the most prevalent and the more challenging NS class because of its lower magnitudes. The data was recalibrated using the N_1 class and a walking algorithm[51] . The m/z match tolerance was set to 1 *ppm*. For the N_1 class the results demonstrated that the reference peak list output by Themis has a similar chemical composition after processing. Plots of contributions by double bond equivalents (DBE) and carbon number for the N_1 class are shown in Figure S6. Figure S6 demonstrates that the assignments were very similar despite the output from Themis containing a fraction of the number of peaks, indicating that information was not being lost during the processing. Themis is expected to improve picking of peaks of low S/N ratio and therefore we next looked at the NS class which forms a smaller contribution to the profile. Figure 6 shows the contributions of homologous series to the NS class, where the NS class included many lower magnitude peaks, as already shown in Figure 5 .

At first glance, a wider range of carbon numbers and DBE appeared to be observed when no processing was used. Closer inspection of the data, however, revealed gaps within the DBE series; this can typically be used to differentiate between likely correct and incorrect assignments within petroleum data, due to the well-known presence of homologous series. The additional assignments in the unprocessed replicates were also associated with higher mass errors, further indicating that they were of questionable validity. Furthermore, manual

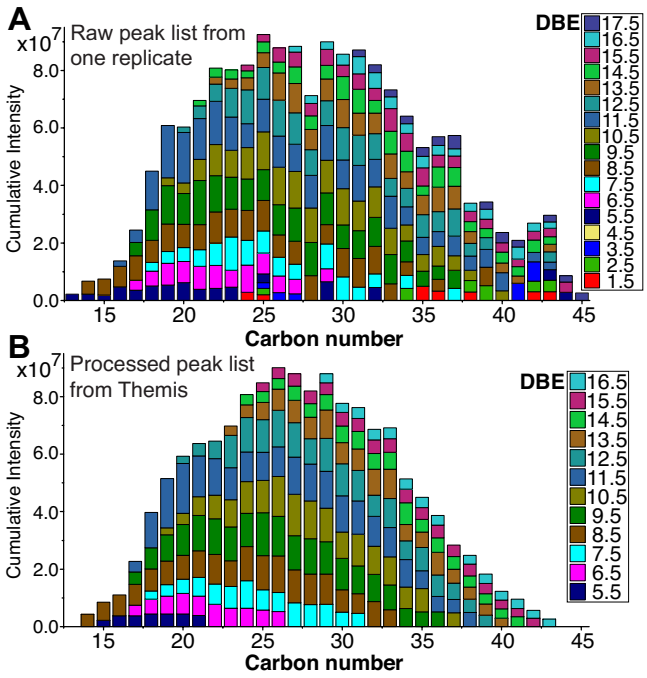

Figure 6: Stacked bar plot of the carbon number and DBE distributions for the NS class for the NIST light sour crude oil sample. The results of the data analyses are shown for: (**A**) a single replicate, where the total peak list (all classes and including noise) comprised approximately 16,400 peaks and for (**B**) the output from Themis using all replicates, where the entire peak list comprised approximately 2,260 peaks.

inspection of the data also revealed that the peaks in question were not consistently observed across the replicates. The combination of these observations provides evidence that the removal of these assignments does not represent a loss of information, but, in fact, a reduction in false positives. After processing with Themis, the series observed were more consistent and the associated range of mass errors was smaller. While Themis reduced the size of the peak list by differentiating noise and inconsistent peaks, information is not being lost. In fact, the processing has facilitated an improvement in data quality by reducing interferences in the analysis from false positives.

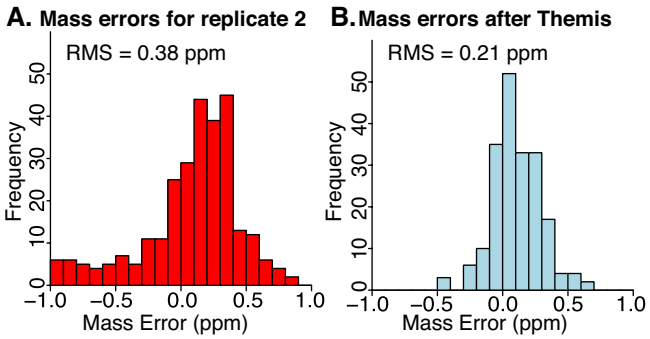

Figure 7: Histogram showing mass errors associated with assignments for the positive ESI mode NIST data for the NS class for one replicate (**A**) and after processing with Themis (**B**)

Figure 7 is a histogram of the mass errors associated with assignments of the NS class for sample A (NIST) before processing (Figure **7.A**) and after Themis processing (Figure **7.B**). The typical mass errors were below 1 *ppm* for both datasets, with a root mean square of 0.38 and 0.21, respectively. The unprocessed replicate displays larger mass errors than data resulting from the processing with Themis as also illustrated by the false positives in Figure **6.A**.

Conclusion

Themis capitalizes on the availability of replicated measurements to generate a single, reliable peak list, while avoiding the *a priori* discarding of low magnitude peaks which typically

occurs when applying signal-to-noise thresholds. At a practical level, the user's workflow is simplified by performing downstream data analysis using a single dataset produced by Themis, instead of working with replicates individually and comparing results at the end. Furthermore, the pre-processing actually led to improved assignment of low magnitude contributions. Dataset sizes and the demand for more reliable, replicated data will increase alongside technological advances in experimental methods. There is an accompanying need to simplify datasets and handle greater numbers of mass spectra. Themis currently performs its tasks within a few minutes and removes the majority of the noise, but there is scope for improvement. For instance, one could incorporate into the analysis peak shape information such as the full width at half the maximum or some chemical prior information to further refine the output reference peak list. In this work it has been found that it is simplistic to use a single parameter threshold, such as S/N ratio, to separate noise from valid peaks; and using m/z in combination with magnitude is a more promising approach. While the application of Themis has been demonstrated using petroleum, it is expected to also be useful for other complex samples. It is intended that Themis will be included in a workflow alongside specialized software for the analysis of different varieties of complex mixtures. The anticipated benefits include faster downstream data analysis, fewer false positives, fewer genuine peaks discarded and hence ultimately an increased confidence in the results of the analysis, which is vital when decision making may be based on the findings.

Acknowledgement

R.G. thanks EPSRC for a PhD studentship through the EPSRC Centre for Doctoral Training in Molecular Analytical Science, grant number EP/L015307/1. D.R. was partially supported by Ramon y Cajal Fellowship RYC-2015-18544 from Ministerio de Economia y Competitividad (Government of Spain). R.G. also thanks David Stranz (Sierra Analytics) for his valuable contributions.

Supporting Information Available

- Figure S1 : Quantile-quantile plots of the magnitudes of the six replicates before and after quantile normalization for the South American crude oil Sample B, negative-ion ESI.
- Figure S2: Quantile-quantile plots of the magnitudes of the five replicates before and after quantile normalization for the South American crude oil Sample B, positive-ion ESI.
- Figure S3: Quantile-quantile plots of the magnitudes of the six replicates before and after quantile normalization for the naphthenic acid Sample C, negative-ion ESI.
- Figure S4: Illustration of the automated detection of outliers and use of a series of replicates to produce an averaged data set for characterization
- Figure S5: Enlargement of the low S/N region between m/z 700 and 710 for the NIST light sour crude oil.
- Figure S6: Stacked bar plot the carbon number and DBE distributions for the N_1 class for the NIST light sour crude oil sample.
- Figure S7: Histograms showing mass errors associated with assignments for the positive ESI mode NIST data for the N_1 class for one replicate and after processing with Themis.
- Figure S8: Multidimensional Scaling (MDS) two-dimensional plot based on the Spearman correlation between magnitudes for each pair of samples before and after Themis

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *25*, 282–283.

(2) Comisarow, M. B.; Marshall, A. G. *Can. J. Chem.* **1974**, *52*, 1997–1999.

(3) Comisarow, M. B.; Marshall, A. G. *Chem. Phys. Lett.* **1974**, *26*, 489–490.

(4) Amster, I. J. *J. Mass Spectrom.* **1996**, *31*, 1325–1337.

(5) Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. *Mass Spectrom. Rev.* **1998**, *17*, 1–35.

(6) Barrow, M. P.; Burkitt, W. I.; Derrick, P. J.; Jennings, K. R.; Dolnikowski, G. G.; Glish, G. L.; Vachet, R. W. *The Analyst* **2005**, *130*, 18.

(7) Schaub, T. M.; Hendrickson, C. L.; Horning, S.; Quinn, J. P.; Senko, M. W.; Marshall, A. G. *Anal. Chem.* **2008**, *80*, 3985–3990.

(8) Qian, K.; Rodgers, R. P.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. *Energy and Fuels* **2001**, *15*, 492–498.

(9) Barrow, M. P.; McDonnell, L. A.; Feng, X.; Walker, J.; Derrick, P. J. *Anal. Chem.* **2003**, *75*, 860–866.

(10) Marshall, A. G.; Rodgers, R. P. *Acc. Chem. Res.* **2004**, *37*, 53–59.

(11) Rodgers, R. P.; Schaub, T. M.; Marshall, A. G. *Anal. Chem.* **2005**, *77*, 20A–27A.

(12) Marshall, A. G.; Rodgers, R. P. *Proc. Natl. Acad. Sci. U.S.A* **2008**, *105*, 18090–18095.

(13) Hsu, C. S.; Hendrickson, C. L.; Rodgers, R. P.; McKenna, A. M.; Marshall, A. G. *J. Mass Spectrom.* **2011**, *46*, 337–343.

(14) Griffiths, M. T.; Da Campo, R.; O'Connor, P. B.; Barrow, M. P. *Anal. Chem.* **2014**, *86*, 527–534.

(15) Rodgers, R. P.; McKenna, A. M. *Anal. Chem.* **2011**, *83*, 4665–4687.

(16) Barrow, M. P.; Witt, M.; Headley, J. V.; Peru, K. M. *Anal. Chem.* **2010**, *82*, 3727–3735.

(17) Headley, J. V.; Barrow, M. P.; Peru, K. M.; Fahlman, B.; Frank, R. A.; Bickerton, G.; McMaster, M. E.; Parrott, J.; Hewitt, L. M. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 1899–1909.

(18) Barrow, M. P.; Peru, K. M.; Headley, J. V. *Anal. Chem.* **2014**, *86*, 8281–8288.

(19) Headley, J. V.; Peru, K. M.; Barrow, M. P. *Mass Spectrom. Rev.* **2016**, *35*, 311–328.

(20) Zhurov, K. O.; Kozhinov, A. N.; Tsybin, Y. O. *Energy and Fuels* **2013**, *27*, 2974–2983.

(21) Headley, J. V.; Peru, K. M.; Janfada, A.; Fahlman, B.; Gu, C.; Hassan, S. *Rapid Commun. Mass Spectrom.* **2011**, *25*, 459–462.

(22) Marshall, A. G.; Hendrickson, C. L. *Annu. Rev. Anal. Chem.* **2008**, *1*, 579–599.

(23) Pomerantz, A. E.; Mullins, O. C.; Paul, G.; Ruzicka, J.; Sanders, M. *Energy and Fuels* **2011**, *25*, 3077–3082.

(24) Smith, E. A.; Park, S.; Klein, A. T.; Lee, Y. J. *Energy and Fuels* **2012**, *26*, 3796–3802.

(25) Dunning, H. N.; Moore, J. W.; Bieber, H.; Williams, R. B. *J. Chem. Eng. Data* **2000**, *497*, 546–549.

(26) Baker, E. W.; Yen, T. F.; Dickie, J. P.; Rhodes, R. E.; Clark, L. F. *J. Am. Chem. Soc.* **1967**, *89*, 3631–3639.

(27) Barrow, M. P. *Biofuels* **2010**, *1*, 651–655.

(28) Cho, Y.; Ahmed, A.; Islam, A.; Kim, S. *Mass Spectrom. Rev.* **2015**, *34*, 248–263.

(29) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A* **2000**, *97*, 10313–10317.

(30) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320–332.

(31) Kaur, P.; O'Connor, P. B. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 459–468.

(32) Du, P.; Kibbe, W. A.; Lin, S. M. *Bioinformatics* **2006**, *22*, 2059–2065.

(33) Mantini, D.; Petrucci, F.; Pieragostino, D.; Del Boccio, P.; Di Nicola, M.; Di Ilio, C.; Federici, G.; Sacchetta, P.; Comani, S.; Urbani, A. *BMC Bioinformatics* **2007**, *8*, 101.

(34) Meuleman, W.; Engwegen, J. Y. M. N.; Gast, M.-C. W.; Wessels, L. F. a.; Reinders, M. J. T. *BMC Bioinformatics* **2009**, *10 Suppl 1*, S51.

(35) Hur, M.; Oh, H. B.; Kim, S. *Bull. Korean Chem. Soc.* **2009**, *30*, 2665–2668.

(36) Zhurov, K. O.; Kozhinov, A. N.; Fornelli, L.; Tsybin, Y. O. *Anal. Chem.* **2014**, *86*, 3308–3316.

(37) Kilgour, D. P. A.; Hughes, S.; Kilgour, S. L.; Mackay, C. L.; Palmblad, M.; Tran, B. Q.; Goo, Y. A.; Ernst, R. K.; Clarke, D. J.; Goodlett, D. R. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 253–262.

(38) Patterson, S. D. *Nature Biotechnol* **2003**, *21*, 221–2.

(39) Chen, L.; Leng Yap, Y. *J. Am. Soc. Mass Spectrom.* **2008**, *19*, 46–54.

(40) Johnson, K. L.; Mason, C. J.; Muddiman, D. C.; Eckel, J. E. *Anal. Chem.* **2004**, *76*, 5097–5103.

(41) McIlwain, S.; Page, D.; Huttlin, E. L.; Sussman, M. R. *Bioinformatics* **2007**, *23*, i328.

(42) Park, K.; Joo, Y. Y.; Lee, S.; Paek, E.; Park, H.; Jung, H. J.; Lee, S. W. *Anal. Chem.* **2008**, *80*, 7294–7303.

(43) Lopez-Fernandez, H.; Santos, H. M.; Capelo, J. L. *BMC Bioinformatics* **2015**, *16*, 1–12.

(44) Mass-Up - mass spectrometry utility for proteomics. `http://sing.ei.uvigo.es/mass-up/`, Accessed: 12/4/2017.

(45) Araújo, J. E.; Santos, T.; Jorge, S.; Pereira, T. M.; Reboiro-Jato, M.; Pavón, R.; Magriço, R.; Teixeira-Costa, F.; Ramos, A.; Santos, H. M. *Anal. Methods* **2015**, *7*, 7467–7473.

(46) Araújo, J. E.; Jorge, S.; Magriço, R.; Costa, T. E.; Ramos, A.; Reboiro-Jato, M.; Riverola, F.; Lodeiro, C.; Capelo, J. L.; Santos, H. M. *Talanta* **2016**, *152*, 364–370.

(47) Santos, T.; Capelo, J. L.; Santos, H. M.; Oliveira, I.; Marinho, C.; Gonçalves, A.; Araújo, J. E.; Poeta, P.; Igrejas, G. *J. Proteomics* **2015**, *127*, 321–331.

(48) Klitzke, C. F.; Corilo, Y. E.; Siek, K.; Binkley, J.; Patrick, J.; Eberlin, M. N. *Energy and Fuels* **2012**, *26*, 5787–5794.

(49) Purcell, J. M.; Merdrignac, I.; Rodgers, R. P.; Marshall, A. G.; Gauthier, T.; Guibard, I. *Energy and Fuels* **2010**, *24*, 2257–2265.

(50) Xian, F.; Hendrickson, C. L.; Blakney, G. T.; Beu, S. C.; Marshall, A. G. *Anal. Chem.* **2010**, *82*, 8807–8812.

(51) Savory, J. J.; Kaiser, N. K.; McKenna, A. M.; Xian, F.; Blakney, G. T.; Rodgers, R. P.; Hendrickson, C. L.; Marshall, A. G. *Anal. Chem.* **2011**, *83*, 1732–1736.

(52) Hur, M.; Yeo, I.; Park, E.; Kim, Y. H.; Yoo, J.; Kim, E.; No, M. H.; Koh, J.; Kim, S. *Anal. Chem.* **2010**, *82*, 211–218.

(53) Hur, M.; Yeo, I.; Kim, E.; No, M.-h.; Koh, J.; Cho, Y. J.; Lee, J. W.; Kim, S. *Energy and Fuels* **2010**, *24*, 5524–5532.

(54) Payne, T. G.; Southam, A. D.; Arvanitis, T. N.; Viant, M. R. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 1087–1095.

(55) Pruski, P.; MacIntyre, D. A.; Lewis, H. V.; Inglese, P.; Correia, G. D. S.; Hansel, T. T.; Bennett, P. R.; Holmes, E.; Takats, Z. *Analytical Chemistry* **2017**, *89*, 1540–1550.

(56) Hampel, F. R. *Source J. Am. Stat. Assoc.* **1974**, *69*, 383–393.

(57) Huber, P. J. *Robust statistics*; New York: John Wiley, 1981.

(58) Rousseeuw, P. J.; Croux, C. *J. Am. Stat. Assoc.* **1993**, *88*, 1273–1283.

(59) Amaratunga, D.; Cabrera, J. *J. Am. Stat. Assoc.* **2001**, *96*, 1161–1170.

(60) Bolstad, B. M.; Irizarry, R. a.; Strand, M. *Bioinformatics* **2003**, *19*, 185–193.

(61) Dempster, A.; Laird, N.; Rubin, D. B. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–38.

(62) Fraley, C.; Raftery, A. E. *J. Am. Stat. Assoc.* **2002**, *97*, 611–631.

(63) Fraley, C.; Raftery, A. E.; Murphy, T. B.; Scrucca, L. *Tech. Rep. 597, Univ. Washingt.* **2012**, 1–50.

(64) Smyth, G. K.; Speed, T. *Methods* **2003**, *31*, 265–273.

(65) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. *Nucleic Acids Research* **2015**, *43*, e47.

(66) Da Campo, R.; Barrow, M. P.; Shepherd, A. G.; Salisbury, M.; Derrick, P. J. *Energy and Fuels* **2009**, *23*, 5544–5549.

(67) Newton, R.; Wernisch, L. *R News* **2007**, *7*, 32–35.

Graphical TOC Entry